THE SCIENTIFIC METHOD IN THE SCIENCE OF MACHINE LEARNING

Jessica Zosa Forde
Project Jupyter
jzf2101@columbia.edu

Michela Paganini
Facebook AI Research
michela@fb.com

ABSTRACT

In the quest to align deep learning with the sciences to address calls for rigor, safety, and interpretability in machine learning systems, this contribution identifies key missing pieces: the stages of hypothesis formulation and testing, as well as statistical and systematic uncertainty estimation – core tenets of the scientific method. This position paper discusses the ways in which contemporary science is conducted in other domains and identifies potentially useful practices. We present a case study from physics and describe how this field has promoted rigor through specific methodological practices, and provide recommendations on how machine learning researchers can adopt these practices into the research ecosystem. We argue that both domain-driven experiments and application-agnostic questions of the inner workings of fundamental building blocks of machine learning models ought to be examined with the tools of the scientific method, to ensure we not only understand effect, but also begin to understand cause, which is the raison d'être of science.

1 INTRODUCTION

As machine learning (ML) matures and ensconces itself in the daily lives of people, calls for assurances of safety, rigor, robustness, and interpretability have come to the forefront of discussion. Many of these concerns, both social and technical, spring from the lack of understanding of the root causes of performance in ML systems (Rahimi & Recht, 2017b; Lipton & Steinhardt, 2018; Sculley et al., 2018b; 2015; Woods, 2018). Researchers have emphasized the need for more thorough examination of published results. Replications of previous work in GANs (Lucic et al., 2018), language modeling (Melis et al., 2018; Vaswani et al., 2017), Bayesian neural networks (Riquelme et al., 2018), reinforcement learning (Henderson et al., 2018; Mania et al., 2018; Nagarajan et al., 2019) that attempted to understand if extensions to simpler models affect performance have yielded negative results. Large scale studies of published work have attempted to provide the community with independent re-implementations of machine learning experiments (Pineau et al., 2017; Tian et al., 2019). At the same time, researchers have demonstrated that trained models fail to generalize to novel scenarios (Recht et al., 2019; Zhang et al., 2019; 2018) and suggested that this decrease in performance may be due to confounders in training data (Stock & Cisse, 2018; Zhang et al., 2017). These inconsistencies in performance are particularly troubling in deployed systems, whose errors have real world implications (Buolamwini & Gebru, 2018; Rudin, 2018; Zech et al., 2018). Many have identified various social factors within the research ecosystem that contribute to these challenges (Lipton & Steinhardt, 2018; Sculley et al., 2018a;b; Henderson & Brunskill, 2018; Gebru et al., 2018; Mitchell et al., 2019).

This position paper identifies ML research practices that differ from the methodologies of other scientific domains. We conjecture that grounding ML research in statistically sound hypothesis testing with careful control of nuisance parameters may encourage the publication of advances that stand the test of time. As in Rahimi & Recht (2017a); Rahimi (2018), we consider methodological techniques in physics to suggest approaches that ML research scientists can adopt to ground their work in the conventions of empirical science. Proper application of the scientific method can help researchers understand factors of variation in experimental outcomes, as well as the dynamics of components in ML models, which would aid in ensuring robust performance in real world systems.

Explorimentation, or the practice of poking around to see what happens, while appropriate for the early stages of research to inform and guide the formulation of a plausible hypothesis, does not constitute sufficient progress to term the effort scientific. As the field evolves towards a more mature state, a pragmatic approach ("It works!") is no longer sufficient. It is necessary to get past the stage of exploratory analysis and start venturing into rigorous hypothesis formulation and testing.

2 THE SCIENTIFIC METHOD

Starting from the assumption that there exists accessible ground truth, the scientific method is a systematic framework for experimentation that allows researchers to make objective statements about phenomena and gain knowledge of the fundamental workings of a system under investigation. The scientific method can also be seen as a social contract, a set of conventions that the community of researchers agrees to follow to everyone's benefit, rather than the one and only path to knowledge. The reasons for its existence are practical. The goal of adopting a universally accepted set of guidelines is to be able to gain confirmable knowledge and insight into patterns of behavior that can be modeled systematically and can be reproduced in an experimental environment, under the assumption of there being a fundamental cause that drives the observed phenomena.

Central to this framework is the formulation of a scientific hypothesis and an expectation that can be falsified through experiments and statistical methods (Jeffrey, 1956; Popper, 1959; Lindberg, 1967; Arbuthnot, 1710; Student, 1908); failure to include these steps is likely to lead to unscientific findings. A well-formed hypotheses generates logical predictions and motivates experimentation: "If the hypothesis is right, then I should expect to observe..." Rather than relying on untested speculations after the experiment has been conducted, both the null and alternative hypotheses ought to be stated prior to collecting data and statistical testing. Examples of hypothesis statements in the literature include those in Nguyen et al. (2015); Azizzadenesheli et al. (2018); Recht et al. (2019).

At the base of scientific research lies the notion that an experimental outcome is a random variable, and that appropriate statistical machinery must be employed to estimate the properties of its distribution. It is important not to forget the role of chance in measurements: "What are the odds of observing this result if the hypothesized trend is not real and the result is simply due to a statistical fluctuation?" The first step towards a scientific formulation of ML then demands a more dramatic shift in priorities from drawing and recording single instances of experimental results to collecting enough data to gain an understanding of population statistics. Since abundant sampling of observations might be prohibitive due to resource constraints, the role of statistical uncertainties accompanying the measurement becomes vital to interpret the result.

Neal (1998); Dietterich (1998); Nadeau & Bengio (2000); Bouckaert & Frank (2004) provided ML researchers with methods of statistical comparison across experimental results on supervised tasks. Demšar (2006) proposes methods for comparing two or more classifiers across multiple datasets, but also observes a lack of consensus in statistical testing methods among ICML papers. These methods are not ubiquitous means of comparison between ML models and may not be applicable to all research questions. Henderson et al. (2018) measure variation in performance in deep reinforcement learning (RL) models across parameter settings, and provides statistical tests to accompany some of these measurements, but they acknowledge that deep RL does not have common practices for statistical testing.

One may argue that the emphasis on hypotheses and statistical testing has created strong incentives in other fields to produce results that show $p \leq 0.05$. These fields face a reproducibility crisis (Ioannidis, 2005; Baker, 2016) and discourage negative results (Rosenthal, 1979; Brodeur et al., 2016; Schuemie et al., 2018). To ensure that these hypotheses are created prior to the start of an experiment, researchers developed systems for pre-registration of hypotheses (Dickersin & Rennie, 2003; Nosek et al., 2018). One of the most notable systems is `clinicaltrials.gov`, which contains information on 300 thousand clinical trials occurring in 208 countries; registration is required for U.S. trials. In at least 165 journals, pre-registered studies may publish results, positive or negative, as a peer-reviewed registered report (Center for Open Science; Nosek & Lakens, 2014; Chambers et al., 2014). A study of 127 registered reports in biomedicine and psychology found that 61% published negative results. Researchers in ML have called for the publication of negative results that are grounded careful experimentation, and such publication systems may contribute to this shift in research practices.

3 CASE STUDY: EXPERIMENTAL HIGH ENERGY PHYSICS

As reference, we offer insight into experimental high energy physics (HEP) — one of many experimental fields in which the scientific method, based on hypothesis formulation and testing, is widely applied. While there has been a long history of statistical physics being utilized in machine learning, such as Ackley et al. (1985) and Geman & Geman (1984), the methodological practices of experimental physics have also been suggested as a model for improved rigor in ML (Rahimi & Recht, 2017a). Platt (1964) specifically identifies HEP as an area of science with notable adherence to hypotheses, a practice he causally links to the rapid advances made possible in that field. Despite the chronological, technical, and methodological differences between ML and HEP, there exist general principles that transcend discipline boundaries and form the necessary components of any rational inquiry that wants to get elevated to the status of science.

In HEP, to check for discovery, the rationale followed to test the validity of a proposed theory (alternative hypothesis) is to compare its observable predictions to those made under a Standard Model-only (Tanabashi et al., 2018) assumption (null hypothesis). Systematic uncertainties enter the estimation of observable effects under both hypotheses, and their careful accounting, along with that of statistical sources of uncertainty, can inform researchers of the expected sensitivity of their analysis before any data is collected. Typically, in HEP, the statistical procedure consists of two steps: the model building phase, and the hypothesis testing phase.

Analytical, parametric models can be derived from first principles or from fits to distributions of observables. The measurement of parameters of interest μ is performed by maximizing the likelihood of the data under the model, while accounting for deviations that can be explained via nuisance parameters ν. Nuisance parameters are allowed to fluctuate within their uncertainties to represent the degree of uncertainty with which their systematic effect on the measurement is known, and provide slack to the fit. They are incorporated in an extended likelihood via multiplicative factors: $L(\mu, \nu) = P_x(x; \mu, \nu)P_z(z; \nu)$, where the measurements of μ and ν occur in statistically orthogonal datasets $x \sim X$ and $z \sim Z$. The portion of the likelihood associated with nuisance parameters estimates the trade-offs associated with moving away from their nominal values in order for the fit to converge. If a nuisance parameter is known with high accuracy, large deviations from its mean will make the solution less probable. High systematic uncertainties reduce the ability to narrow down the plausible values of μ — many will correspond to similarly high values of the likelihood given the freedom to tune the nuisance portion of the model accordingly, which reduces the sensitivity of the experiment. This is the regime in which we expect many ML experiments to live.

A statistical test is run by constructing a suitable test statistic that is solely a function of the value of μ being tested. Manually scanning over values of μ, or relying on asymptotic properties of the test statistic, allows to construct frequentist Neyman confidence intervals (Neyman, 1937) and identify a range of μ values for which the p-value remains lower than a pre-determined magnitude for all values of the nuisance parameters. More details are offered in Appendix A.

3.1 STRICT ANALOGY FOR MACHINE LEARNING EXPERIMENTS

As physicists compute expectations over the number of events in the differential distribution of some discriminative observable, similarly, ML scientists could entertain the idea of investigating neural networks as physical objects and their time evolution as a natural phenomenon that follows the laws of dynamics, modeling ML experimental outcomes as variables, and analyzing them within a HEP-like hypothesis formulation and testing tradition.

For instance, assume one postulated that a new activation function would intervene on information and gradient propagation in a specific, desirable way. First, they would formulate a quantitative, informed hypothesis of the expected behavior of a model with and without the suggested change. Then, if not already available from prior measurements, they would record outcomes from a variety of baseline models on a variety of reasonable datasets of similar characteristics in statistically orthogonal scenarios from the ones in which the measurement is being made, so as to constrain the nuisance parameters. Next, following a simplified additive model, they could assume, for example, that the observed performance using the proposed architectural modification may be separated as follows: (baseline performance) + $\mu\times$ (expected performance improvement from new activation), where both terms on the right hand side can be affected by nuisance parameters. Examples

of nuisance parameters in ML experiments include: the choice of dataset, optimizer, initialization, hyperparameters, activation, normalization, regularization. Their estimation and modeling is not always an exact science and judgment calls will happen (Sinervo, 2003); before the field converges to well-calibrated, agreed-upon uncertainty modeling procedures, coarse and conservative decisions can suffice. However, without appropriate effort from the community in handling and reducing systematic uncertainties from hyperparameter and experimental setup variations, it is unlikely for any analysis to have any sensitivity to the parameter of interest μ. Indeed, since many current ML publications omit this fundamental step, it is plausible that a significant percentage of published work claiming state-of-the-art performance actually has no statistical sensitivity to measure their improvement over competing methods. Indeed, one would expect a large range of values of μ, including 0, to be compatible with a high likelihood value, given the poor constraints on systematics. While reproducibility seeks to ensure "the ability of an experiment to be repeated with minor differences from the original experiment, while achieving the same qualitative results" (Nagarajan et al., 2019), this goal on its own does not explicitly measure the systematic uncertainty of experimental results.

In addition, borrowing another analogy from the field of physics, given the infancy of machine learning as a science, we would like to dissuade researchers from feeling the need to immediately have to come up with and test an all-encompassing "Theory of Everything," and instead focus on measurements of first order behaviors first, as done in Shallue et al. (2018). In loose analogy to perturbation theory in physics, one can study phenomena at different orders in an approximate expansion, starting from main trends and low order effects, then expanding to include higher order interaction terms. In other words, we first need to have a sense of how, say, BatchNorm (Ioffe & Szegedy, 2015) roughly affects convergence, then model the interaction of it with depth, width, and skip connections (He et al., 2016), among others.

4 Conclusions and Recommendations

Hypothesis formulation prior to experimentation and statistical testing are two central pillars of the scientific method which are extremely rarely explicitly found in contemporary machine learning research papers. We strongly suggest that researchers incorporate these stages of experimentation into their work, perhaps by drawing inspiration, when possible, from the methodologies devised by other scientific disciplines. While rejecting mere scientism, we argue that it is necessary for the field to adopt the methodological tooling of empiricism and naturalism by operating in controlled, reproducible, and verifiable settings. While these practices are primarily targeted to researchers interested in fundamental science of deep learning, applied researchers and research engineers will also benefit from the birth of a more principled, scientific subfield of machine learning.

Workshops, conferences, and panels should make the effort to include scientists, philosophers of science, and historians of science in conversations around the necessary steps to favor the transition of deep learning to a science.

In addition, in the spirit of the Workshop on Negative Results in Computer Vision at CVPR 2017, we support the proposal of a future workshop which uses a registered reports model (Center for Open Science). Testable hypotheses would be submitted for approval ahead of time, and resulting contributions will be accepted regardless of their results with respect to accepting or rejecting the null hypothesis, granted that reviewers deem the authors' methods to be technically and scientifically sound.

Finally, we invite the community of reviewers to pay closer attention to the accounting of statistical and systematic uncertainties which plague many state-of-the-art results, and consider the scientific robustness of claims. Submissions should not be discouraged for conflicting with other results, as long as prior art is acknowledged and confronted, and the authors are explicit about their result being incompatible with other findings. As Gauch & Gauch Jr (2003) state, citing Megill (1994),"objective truth expresses beliefs 'towards which all inquirers of good will are destined to converge.'"

References

David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for boltzmann machines. *Cogn. Sci.*, 9(1):147–169, January 1985. URL `http://www.sciencedirect.com/science/article/pii/S0364021385800124`.

John Arbuthnot. An argument for divine providence, taken from the constant regularity observ'd in the births of both sexes. *Philosophical Transactions of the Royal Society of London*, 27 (328):186–190, 1710. URL https://royalsocietypublishing.org/doi/abs/10.1098/rstl.1710.0011.

Kamyar Azizzadenesheli, Brandon Yang, Weitang Liu, Emma Brunskill, Zachary C Lipton, and Animashree Anandkumar. Surprising negative results for generative adversarial tree search. In *Critiquing and Correcting Trends in Machine Learning Workshop, NeurIPS*, June 2018. URL http://arxiv.org/abs/1806.05780.

Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452–454, May 2016. URL http://dx.doi.org/10.1038/533452a.

Remco R Bouckaert and Eibe Frank. Evaluating the replicability of significance tests for comparing learning algorithms. In *Advances in Knowledge Discovery and Data Mining*, pp. 3–12. Springer Berlin Heidelberg, 2004. URL http://dx.doi.org/10.1007/978-3-540-24775-3_3.

Abel Brodeur, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. Star wars: The empirics strike back. *Am. Econ. J. Appl. Econ.*, 8(1):1–32, January 2016. URL https://www.aeaweb.org/articles?id=10.1257/app.20150044.

Andy Buckley, Jonathan Butterworth, Leif Lonnblad, David Grellscheid, Hendrik Hoeth, James Monk, Holger Schulz, and Frank Siegert. Rivet user manual. *Comput. Phys. Commun.*, 184: 2803–2819, 2013. doi: 10.1016/j.cpc.2013.05.021.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A Friedler and Christo Wilson (eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 77–91, New York, NY, USA, 2018. PMLR. URL http://proceedings.mlr.press/v81/buolamwini18a.html.

Center for Open Science. Registered reports. URL https://cos.io/rr/. Accessed: 2019-3-2.

Christopher D Chambers, Eva Feredoes, Suresh Daniel Muthukumaraswamy, and Peter Etchells. Instead of" playing the game" it is time to change the rules: Registered reports at AIMS neuroscience and beyond. *AIMS Neuroscience*, 1(1):4–17, 2014. URL http://orca.cf.ac.uk/id/eprint/59475.

Jake Cowton, Sunje Dallmeier-Tiessen, Pamfilos Fokianos, L Rueda, P Herterich, J Kunčar, T Šimko, and Tim Smith. Open data and data analysis preservation services for lhc experiments. In *Journal of Physics: Conference Series*, volume 664, pp. 032030. IOP Publishing, 2015.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7(Jan):1–30, 2006. URL http://www.jmlr.org/papers/volume7/demsar06a/demsar06a.pdf.

Kay Dickersin and Drummond Rennie. Registering clinical trials. *JAMA*, 290(4):516–523, July 2003. URL http://dx.doi.org/10.1001/jama.290.4.516.

T G Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.*, 10(7):1895–1923, September 1998. URL https://www.ncbi.nlm.nih.gov/pubmed/9744903.

Hugh G Gauch and Hugh G Gauch Jr. *Scientific method in practice*. Cambridge University Press, 2003.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé Iii, and Kate Crawford. Datasheets for datasets. March 2018. URL http://arxiv.org/abs/1803.09010.

S Geman and D Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(6):721–741, June 1984. URL https://www.ncbi.nlm.nih.gov/pubmed/22499653.

Eilam Gross and Ofer Vitells. Trial factors for the look elsewhere effect in high energy physics. *The European Physical Journal C*, 70(1):525–530, Nov 2010. ISSN 1434-6052. doi: 10.1140/epjc/s10052-010-1470-8. URL https://doi.org/10.1140/epjc/s10052-010-1470-8.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Peter Henderson and Emma Brunskill. Distilling information from a flood: A possibility for the use of Meta-Analysis and systematic review in machine learning research. December 2018. URL http://arxiv.org/abs/1812.01074.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Thirty-Second AAAI Conference on Artificial Intelligence*, April 2018. URL https://aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewPaper/16669.

John P A Ioannidis. Why most published research findings are false. *PLoS Med.*, 2(8):e124, August 2005. URL http://dx.doi.org/10.1371/journal.pmed.0020124.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Richard C Jeffrey. Valuation and acceptance of scientific hypotheses. *Philos. Sci.*, 23(3):237–246, July 1956. URL https://doi.org/10.1086/287489.

D C Lindberg. Alhazen's theory of vision and its reception in the west. *Isis*, 58(3):321–341, 1967. URL https://www.ncbi.nlm.nih.gov/pubmed/4867472.

Zachary C Lipton and Jacob Steinhardt. Troubling trends in machine learning scholarship. In *ICML 2018: The Debates*, July 2018. URL http://arxiv.org/abs/1807.03341.

Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are GANs created equal? a Large-Scale study. In *NeurIPS*, 2018. URL http://arxiv.org/abs/1711.10337.

Eamonn Maguire, Lukas Heinrich, and Graeme Watt. HEPData: a repository for high energy physics data. *J. Phys. Conf. Ser.*, 898(10):102006, 2017. doi: 10.1088/1742-6596/898/10/102006.

Horia Mania, Aurelia Guy, and Benjamin Recht. Simple random search provides a competitive approach to reinforcement learning. In *NeurIPS*. arxiv.org, 2018. URL http://arxiv.org/abs/1803.07055.

Allan Megill. *Rethinking Objectivity*. Duke University Press, 1994.

Gábor Melis, Chris Dyer, and Phil Blunsom. On the state of the art of evaluation in neural language models. In *ICLR*, 2018. URL https://openreview.net/pdf?id=ByJHuTgA-.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229. ACM, January 2019. URL https://dl.acm.org/citation.cfm?doid=3287560.3287596.

Claude Nadeau and Yoshua Bengio. Inference for the generalization error. In S A Solla, T K Leen, and K Müller (eds.), *Advances in Neural Information Processing Systems 12*, pp. 307–313. MIT Press, 2000. URL http://papers.nips.cc/paper/1661-inference-for-the-generalization-error.pdf.

Prabhat Nagarajan, Garrett Warnell, and Peter Stone. Deterministic implementations for reproducibility in deep reinforcement learning. In *AAAI Workshop on Reproducible AI*, 2019. URL http://arxiv.org/abs/1809.05676.

Radford M Neal. Assessing relevance determination methods using DELVE. *Nato Asi Series F Computer And Systems Sciences*, 168:97–132, 1998. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.55.6550&rep=rep1&type=pdf.

J. Neyman. Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. *Phil. Trans. Roy. Soc. Lond.*, A236(767):333–380, 1937. doi: 10.1098/rsta.1937.0005.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 427–436, 2015. URL https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Nguyen_Deep_Neural_Networks_2015_CVPR_paper.pdf.

Brian A Nosek and Daniël Lakens. Registered reports. *Soc. Psychol.*, 45(3):137–141, May 2014. URL https://doi.org/10.1027/1864-9335/a000192.

Brian A Nosek, Charles R Ebersole, Alexander C DeHaven, and David T Mellor. The preregistration revolution. *Proc. Natl. Acad. Sci. U. S. A.*, 115(11):2600–2606, March 2018. URL https://www.pnas.org/content/115/11/2600.

Joelle Pineau, Genevieve Fried, Rosemary Nan Ke, and Hugo Larochelle. ICLR 2018 reproducibility challenge. https://www.cs.mcgill.ca/~jpineau/ICLR2018-ReproducibilityChallenge.html, 2017. URL https://www.cs.mcgill.ca/~jpineau/ICLR2018-ReproducibilityChallenge.html. Accessed: 2018-6-10.

J R Platt. Strong inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science*, 146(3642):347–353, October 1964. URL http://dx.doi.org/10.1126/science.146.3642.347.

Karl Popper. *The Logic of Scientific Discovery*. Routledge, 1959. URL https://philpapers.org/rec/POPTLO-7.

Ali Rahimi. Lessons from optics, the other deep learning. http://www.argmin.net/2018/01/25/optics/, January 2018. URL http://www.argmin.net/2018/01/25/optics/. Accessed: 2018-10-30.

Ali Rahimi and Ben Recht. An addendum to alchemy. http://benjamin-recht.github.io/2017/12/11/alchemy-addendum/, December 2017a. URL http://benjamin-recht.github.io/2017/12/11/alchemy-addendum/. Accessed: 2018-9-30.

Ali Rahimi and Benjamin Recht. Reflections on random kitchen sinks. http://www.argmin.net/2017/12/05/kitchen-sinks/, December 2017b. URL http://www.argmin.net/2017/12/05/kitchen-sinks/. Accessed: 2018-10-30.

Alexander L. Read. Presentation of search results: The CL(s) technique. *J. Phys.*, G28:2693–2704, 2002. doi: 10.1088/0954-3899/28/10/313. [,11(2002)].

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? February 2019. URL http://arxiv.org/abs/1902.10811.

Carlos Riquelme, George Tucker, and Jasper Snoek. Deep bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling. In *ICLR 2018*, February 2018. URL https://openreview.net/forum?id=SyYe6k-CW.

Robert Rosenthal. The file drawer problem and tolerance for null results. *Psychol. Bull.*, 86(3):638, 1979. URL https://psycnet.apa.org/journals/bul/86/3/638/.

Cynthia Rudin. Please stop explaining black box models for high stakes decisions. In *Critiquing and Correcting Trends in Machine Learning, NeurIPS Workshop*, 2018. URL http://arxiv.org/abs/1811.10154.

Martijn J Schuemie, Patrick B Ryan, George Hripcsak, David Madigan, and Marc A Suchard. Improving reproducibility by using high-throughput observational studies with empirical calibration. *Philos. Trans. A Math. Phys. Eng. Sci.*, 376(2128), September 2018. URL http://dx.doi.org/10.1098/rsta.2017.0356.

D Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. Hidden technical debt in machine learning systems. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pp. 2503–2511, Cambridge, MA, USA, 2015. MIT Press. URL http://dl.acm.org/citation.cfm?id=2969442.2969519.

D Sculley, Jasper Snoek, and Alex Wiltschko. Avoiding a tragedy of the commons in the peer review process. In *Critiquing and Correcting Trends in Machine Learning Workshop, NeurIPS*, December 2018a. URL http://arxiv.org/abs/1901.06246.

D Sculley, Jasper Snoek, Alex Wiltschko, and Ali Rahimi. Winner's curse? on pace, progress, and empirical rigor. In *ICLR 2018 Workshop*, February 2018b. URL https://openreview.net/forum?id=rJWF0Fywf.

Christopher J Shallue, Jaehoon Lee, Joseph Antognini, Jascha Sohl-Dickstein, Roy Frostig, and George E Dahl. Measuring the effects of data parallelism on neural network training. November 2018. URL http://arxiv.org/abs/1811.03600.

Pekka K Sinervo. Definition and treatment of systematic uncertainties in high energy physics and astrophysics. *Statistical Problems in Particle Physics, Astrophysics, and Cosmology*, pp. 122–129, 2003.

Pierre Stock and Moustapha Cisse. ConvNets and ImageNet beyond accuracy: Understanding mistakes and uncovering biases. In *ECCV*, 2018. URL http://arxiv.org/abs/1711.11443.

Student. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908. URL http://www.jstor.org/stable/2331554.

M. Tanabashi et al. Review of particle physics. *Phys. Rev. D*, 98:030001, Aug 2018. doi: 10.1103/PhysRevD.98.030001. URL https://link.aps.org/doi/10.1103/PhysRevD.98.030001.

Yuandong Tian, Jerry Ma, Qucheng Gong, Shubho Sengupta, Zhuoyuan Chen, James Pinkerton, and C Lawrence Zitnick. ELF OpenGo: An analysis and open reimplementation of AlphaZero. *arXiv [cs.AI]*, February 2019. URL https://arxiv.org/abs/1902.04522.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł Ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Wouter Verkerke and David Kirkby. The roofit toolkit for data modeling. In *Statistical Problems in Particle Physics, Astrophysics and Cosmology*, pp. 186–189. World Scientific, 2006.

Samuel S Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.

Bronwyn Woods. Expanding search in the space of empirical ML. In *Critiquing and Correcting Trends in Machine Learning Workshop, NeurIPS*, December 2018. URL http://arxiv.org/abs/1812.01495.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Med.*, 15(11):e1002683, November 2018. URL http://dx.doi.org/10.1371/journal.pmed.1002683.

Amy Zhang, Yuxin Wu, and Joelle Pineau. Natural environment benchmarks for reinforcement learning. November 2018. URL http://arxiv.org/abs/1811.06032.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations*, 2017. URL https://openreview.net/pdf?id=Sy8gdB9xx.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, and Yoram Singer. Identity crisis: Memorization and generalization under extreme overparameterization. February 2019. URL http://arxiv.org/abs/1902.04698.

A Additional Information on Hypothesis Testing in HEP

A commonly adopted statistic in HEP is the profile likelihood-ratio $\lambda(\mu) = \frac{L(\mu,\hat{\hat{\nu}}(\mu))}{L(\hat{\mu},\hat{\nu})} = \frac{\max_\nu L(\mu,\nu)}{L(\hat{\mu},\hat{\nu})}$, in which nuisance parameters ν are profiled, *i.e.* at the numerator, they are first expressed as functions of μ and maximized to remove any dependence on them ($\hat{\hat{\nu}}(\mu)$ is the conditional maximum likelihood estimator of a nuisance parameter), while the denominator represents the maximum of the unconstrained likelihood and is jointly maximized over μ and ν (the single-hat notation indicates the maximum likelihood estimator of a parameter).

The p-value

$$p_\mu = \int_{t_{\mu,\ \text{obs.}}}^{\infty} f(t_\mu|\mu)dt_\mu \tag{1}$$

can be expressed in terms of the reparametrized profile likelihood ratio $t_\mu = -2\ln\lambda(\mu)$, which enjoys better numerical stability. In the asymptotic limit, $f(t_\mu)$ is χ^2 distributed (Wilks, 1938).

Incidentally, HEP adopts a unique approach to persistency of analyses: analysis preservation and combination are primarily achieved by sharing serialized versions of the likelihood model (Verkerke & Kirkby, 2006), though solutions for analysis code sharing and data preservation have recently gained more traction in the community (Buckley et al., 2013; Maguire et al., 2017; Cowton et al., 2015).

We refer interested readers to the extensive prior literature for important considerations and information about advanced statistical methods and topics commonly used in HEP, such as the CLs method (Read, 2002) (where the probability for the signal + background hypothesis is normalized by the background-only probability to avoid spurious exclusions of the null hypothesis in low experimental sensitivity analyses) or the "Look-Elsewhere Effect" (Gross & Vitells, 2010) (which points to the need for accounting for large parameter spaces and numbers of trials when assessing the true statistical significance of an observation).